EQUIP2
Educational Quality Improvement Program
Policy ▪ Systems ▪ Management

ISSUES BRIEF

Measuring Learning Outcomes in Developing Countries: A Primer

Why Measure Learning Outcomes?

Educational quality can no longer be measured solely with inputs or simple outputs (UNESCO 1990, 2000). While it is important to know how much is being spent on education, the number of qualified teachers, or the percentage of an age cohort that reaches a certain level of education, it is just as important to know what students have learned and can do. Ensuring that educational systems are producing students with the knowledge, skills, and attitudes they need for personal development, productive lives, and citizenship is imperative. This is particularly true in developing countries that want to join the global marketplace and improve the quality of life of their people.

Measuring learning outcomes provides useful information for improving educational planning, management, and teaching. Its importance is underscored by initiatives such as Education for All, which requires that countries receiving assistance improve the measurement of learning outcomes and the systems used to regularly monitor education.

The measurement of learning outcomes starts in classrooms, where teachers informally evaluate students' knowledge and performance. As students progress through the system, they may be required to take more official tests. These are used to sort students within schools, award certifications, or meet requirements for entering higher levels of education or particular schools.

Learning outcomes are also measured at subnational and national levels. Measurements may benchmark where the school or system is, allowing comparisons to similar systems or points in time. The measurements may also be used to make decisions about the allocation of resources or hold officials responsible.

A Few Words on Terminology

The terminology related to learning outcomes measurement is not always applied precisely. In this primer, *measuring learning outcomes* is an umbrella term, and the terms *evaluation of students* and *outcomes* are used in similarly generic ways, conveying that judgments are being made about what students have learned.

Assessment generally refers to system-level activities. *Examination* refers to individual-level activities. *Test* frequently refers to individual-level activities—to an instrument rather than an overall program.

Monitoring is the periodic assessment of an education system, whether of a nation, province, or school. *Benchmarking* is the assessed appraisal of an education system by comparing it to some other system or systems. *Accountability* refers the use of assessment results to make a value judgment or to hold someone responsible for the quality of output, such as a teacher for a class or a principal for a school.

EQUIP2 is funded by the U. S. Agency for International Development

Cooperative Agreement No. GDG-A-00-03-00008-00

A Framework for Understanding the Measurement of Learning Outcomes

The measurement of learning outcomes can be grouped into four categories relating to who or what is being measured and the purpose of the measurement (Table 1):

- *Low-stakes measurement of individual outcomes* is the evaluation of individual students by classroom teachers or others for the purpose of understanding or affirming students' knowledge and abilities and informing teaching practice. This category includes what is referred to as continuous assessment. It also includes tests of curricular mastery and diagnostic tests that identify students' entry-level knowledge and skills. The evaluations may be less formal (drawing on verbal questioning by a teacher) or more formal (influenced by written tests for which grades or marks are assigned).
- *High-stakes measurement of individual outcomes* is evaluation with major consequences that relate to individual students' knowledge and performance. This category includes entrance examinations for particular schools or levels of education and tests whose results sort students within schools or tracks. Other high-stakes tests certify the completion of a program or eligibility for the next educational level.
- *Low-stakes measurement of system outcomes* is evaluation of students' knowledge and performance as a group (within a school, region, or nation). The purpose is to monitor the group over time or compare or benchmark its performance to similar groups.
- *High-stakes measurement of system outcomes* is evaluation of students' knowledge and performance as a group that are used to hold teachers, principals, and education officials accountable. Consequences can include changes in allocated resources or staffing, discontinued operation, or transfers. The consequences are lesser for assessments made for public accountability, such as when school-level results are published in league tables or school report cards.

Table 1. Measurement of Learning Outcomes: Four Categories

Who or what is being measured?	**What is the purpose of the measurement?**	
	Low stakes	*High stakes*
Individual	Continuous assessment and tests of curriculum mastery	Examinations for entrance, sorting, and certification
System (national, provincial, school)	Assessment for monitoring and benchmarking	Assessment for accountability

At the individual level, conceptual differences between low- and high-stakes measurement may range from informal classroom activities to major system-wide examinations for explicit purposes. At the system level, conceptual differences are not so great: they relate mainly to the use to which the information is put rather than the scope of the information gathered.

Measuring Learning Outcomes at the System Level

This section provides information on key considerations in building and administering a valid and meaningful assessment of learning outcomes at the system level. Learning outcomes refers to both cognitive and affective outcomes (skills and attitudes). Test instruments collect data on students' knowledge and skills while questionnaires collect data on important contextual variables relating to students, schools, and their families. Both are analyzed in relation to performance or as outcomes themselves.

Quality assessment programs are marked by certain common features.

1. *Any assessment should be based on a sound framework of what is to be measured.*

- The framework is the foundation on which an assessment is built.
- It needs to define explicitly what is to be measured (the parameters of the domain), what tasks represent the skills and abilities in the domain and how they can be organized, and the key characteristics of those tasks.
- The framework provides the outline for test development and the rationale behind that outline. The framework is thus the vehicle for deciding at a detailed level the kinds of evidence that need to be collected. It is also a mechanism for building consensus (Kirsch 2003).

2. *Any assessment program should strive to be valid, reliable, and fair.*

- Assessments need to be valid in three ways. To be valid in terms of *content* means that what is being measured is representative of the curriculum, standards, or framework underlying the assessment. To be valid in terms of *construct* means ensuring that the instrument measures what it says it is measuring. For an assessment to be valid *on its face*, key stakeholders perceive it as a worthy measurement.
- A reliable assessment means that test instruments, sampling plans, implementation specifications, and quality control mechanisms would produce the same results if the assessment were immediately replicated with another sample of students.
- A fair assessment is perceived and empirically validated as such. Its test instruments do not discriminate against subgroups of students by such characteristics as gender or rural status. Further, in a fair assessment, what is measured is appropriate for the students and domain assessed.

3. *Any assessment program should carefully consider how results are presented.*

- Results presented must have been analyzed in statistically sound ways, taking into account technical standards, statistical significance, and minimum response rates.
- Formats used to present data should match audiences and program purposes. This is important because assessments can serve a variety of purposes. They can inform policy, increase public awareness, and stimulate debate. They can also be used to monitor equity, help identify a system's strengths and weaknesses, assist teachers' efforts to raise achievement, and help principals and teachers understand school-level results and performance in particular content areas.

Designing Assessments

Most design decisions are directly influenced by the purposes of the assessment and the ways its information will be used. A clear conception of both is absolutely imperative.

Applying the Framework in the U.S. Context

Low-stakes measurement of individual outcomes
In this category are tests, quizzes, written assignments, discussions, and observations that inform a classroom teacher about a student's progress, influence the curriculum received by a student in the classroom, or lead to a report card grade.

High-stakes measurement of individual outcomes
In this category are tests that affect decisions about a student's educational future, including school entrance exams, exams certifying a level of achievement, tests that determine placement within a school, and college-entry exams.

Low-stakes measurement of system outcomes
This category includes the Nation's Report Card—the National Assessment of Educational Progress (NAEP)—that monitors students' performance at national and state levels and over time. It also includes the United States' participation in international assessments.

[Continued on next page]

High-stakes measurement of system outcomes
This category includes measures prescribed by "No Child Left Behind," which ties federal funding to states' implementation of statewide standards and aligned assessments in key subjects and grades. No Child Left Behind allows penalties to be assigned for districts and schools failing to make yearly progress toward these standards and gives parents the right to transfer their child to another school if the current school is consistently deemed to be failing.

Identifying Who Is Assessed

The first task is identifying the target population about whom information is needed. Usually this means identifying the grade or grades of students to be assessed. This identification includes deciding if any groups of students are to be excluded and, if so, on what grounds (such as language spoken or special education status). Assessment designers must balance desire for representativeness and inclusiveness against fairness to individual students and the feasibility of testing them.

Identifying who is assessed also requires decisions about whether to sample students. High-quality assessment programs offer detailed plans for sampling design, taking into account the eventual uses of the data and identifying the level they will be presented.

Identifying What Is Measured

The second task in assessment design is identifying what is to be measured. This means the domain or domains measured—perhaps reading, mathematics, or motivation for learning. Some assessments focus on a discrete number of domains while others seek information on a wide range. For example, Ireland's national assessment focuses on English reading, mathematics, and Irish language, while New Zealand's National Education Monitoring Project collects data on nearly every curricular area. Whatever the approach, the strategy makes explicit the rationale for selecting domains, offers explanatory definitions, and describes how the domain will be organized for measurement.

Some assessment frameworks focus on the curriculum. For example, England's national testing program assesses students' progress in attaining the national curriculum at three key educational stages. Alternatively, a framework may be organized more normatively, concentrating on what should be learned within a domain. A panel of experts and stakeholders decide what NAEP should measure for various subject areas and grades. This approach is practical for the highly decentralized educational system in the United States.

Multiple Choice or Open-Ended Questions

The framework should indicate the types of items on which the assessment will rely. They may be multiple-choice (where students select an answer from a list), or open-ended (where students craft their own written responses). Open-ended responses may be either short and objective—such as an addition problem—or long and more subjective. Most assessments rely on a mix of multiple-choice and open-ended items, though the proportion varies significantly.

Open-ended items are often considered more authentic than multiple-choice items because they are more similar to everyday problems and situations. However, open-ended items may cause difficulty if students are not used to the format. They can also be significantly more expensive to mark, since they require subjective analysis by individuals trained in rating responses. Some research in OECD countries suggests gender differences in performance caused by item format. This should be taken into consideration to avoid constructing a biased assessment (Bolger and Kellaghan 1990).

Questions of Timing

Other important considerations relate to how often assessments are conducted and how frequently their data are needed. When multiple domains are assessed, the question is whether they are assessed the same year with the same students—to analyze correlation of performance—or assessed in different years.

Methodologies and Who Administers Them

Statistical methodologies are another important consideration. Recent advances in item-response theory (IRT) permits reporting by proficiency levels. These are used to predict whether students are likely to respond correctly. NAEP and Canada's School Achievement Indicators Program now report on the percentage of students reaching graduated levels of proficiency in domains. Canada combines this with a criterion-based approach, setting expectations for the percentage of students who will reach certain levels of proficiency and comparing actual and expected percentages.

Who is involved in delivering, marking, and analyzing assessment results is a final consideration. The assessment's design and implementation may be contracted to national research centers or private organizations with relevant expertise. In other cases, the assessment is conducted within an education ministry, perhaps with assistance from outside consultants. In either case, teachers, curriculum specialists, or university researchers are usually involved in creating frameworks and instruments.

Another question is who administers the tests within schools, marks open-ended portions, analyses data, and produces reports and products. Each role requires careful thought. Assessment staff are arguably more neutral in administering tests than classroom teachers or principals. All markers, whether from the private sector, university, or classroom, need training to ensure reliability.

Who Does What

The overall strategy of the framework guiding, defining, and describing an assessment program is usually led and approved by policymakers and key stakeholders. The frameworks themselves may be developed by groups of subject-matter and testing experts. The expert groups generally construct the tests: they find or develop test items that cover the domain, which is defined in the framework. Many more items are needed for the initial pool than will be used in the final test. Some will turn out not to function well.

Assessment instruments and supplementary questionnaires are field tested, usually at least one year before implementation. Field testing allows designers to choose the best functioning items, make revisions, and ensure no systematic differences exist in the way items function, for example, between boys and girls. Field testing also verifies that most students can complete the test and supplementary questionnaires within the time allotted.

Most assessments are sample-based rather than universal. They aim to be representative of the national student population. Depending on data uses, this may require more than a simple random sample, especially if data are disaggregated to subnational or school levels or by student subgroups. For example, certain subgroups—such as private school students—may be oversampled to permit later comparisons with other subgroups.

The overall strategy of the framework guiding, defining, and describing an assessment program is usually led and approved by policymakers and key stakeholders. The frameworks themselves may be developed by groups of subject-matter and testing experts.

Participation in international and regional assessments allows countries to benchmark the knowledge and performance of their students. This is another key reason for undertaking system-level assessment.

Most assessment programs have agreed-upon sampling plans. These indicate the sampling frame used, how students or schools will be sampled, and how sampling will affect later analysis. The plans also describe how replacement units are selected as well as acceptable response rates for voluntary assessments. Sampling plans also address exclusion issues such as disability or immigrant status.

Past international assessments have recommended response rates of at least 85 percent of schools and 90 percent of students, with no more than 10 percent of students initially excluded. In international assessment efforts, an international sampling referee aims to ensure that countries' sampling plans result in comparable populations being assessed.

Most assessment programs take advantage of recent methodological advances, including IRT and matrix sampling. In the latter case, giving students differing test booklets increases assessment items while allowing performance to be compared with other students for the overall domain.

Though how results are reported and to whom may vary widely, national reports usually present overall mean scores, distribution of scores (from low to high percentiles of performance), percentages of students reaching proficiency levels or other norm- or criterion-referenced benchmarks, and breakdowns by subgroups.

Despite pressures to make the full assessment public, often only about half of its items can be disclosed. This allows a balance between educating the public about what the assessment measures and maintaining valid trend measures, because many assessments aim to measure trends in performance in a certain grade and subject over time. While it is possible to make items public and replace them with new ones with similar characteristics, those kept confidential can be reused. This ensures that the overall measure stays the same and students do not receive unfair advantage from seeing released items.

Assessment programs also generally provide for quality control mechanisms, including translation checking, marker training, and data cleaning. The mechanisms include written procedures for implementation and random checks to ensure they are carried out. Typically, quality control monitors are sent to a sample of the schools to observe practices and provide technical support.

International Assessment Activities

Participation in international and regional assessments allows countries to benchmark the knowledge and performance of their students. This is another key reason for undertaking system-level assessment. Benchmarking can also occur with national-level assessments if intranational comparisons are made.) Three major international student assessments that provide benchmarking opportunities for developing countries are described below. An assessment relating to the overall learning outcomes of the adult population is also referenced:

- The Progress in Reading Literacy Study (PIRLS) of the International Association for the Evaluation of Educational Achievement (IEA) is designed to measure trends in the reading literacy achievement of 4th-grade students. It concentrates on factors at home and in school that facilitate acquisition of literacy in young children. Thirty-five countries (including 18 non-OECD countries) participated in the first cycle of PIRLS in 2001, and enrollment is open for the second cycle in 2006. PIRLS views reading as an interactive, constructive process, emphasizing the importance of students' abilities to use it for different purposes. This is not dissimilar to the view of reading literacy in the OECD assessment of older students, described below.

- The IEA's Trends in International Mathematics and Science Study (TIMSS) collects data on mathematics and science achievements of 4th- and 8th-grade students as well as background factors relating to achievement. Its frameworks are grounded in common curricular and content areas. When TIMSS was administered in 1995, 45 countries participated. It was then the largest assessment of student performance undertaken. It since has been repeated: in 1999 with 38 countries, and 2003 with 51 countries (including 35 non-OECD countries). Plans are underway for a fourth cycle in 2007.

- The OECD's Program for International Student Assessment (PISA) collects data every three years on 15-year-old students, as well as background information on schools and students. The data relate to students' reading, mathematical, ICT, and scientific literacy, and their competencies in such areas as problem solving. PISA was administered in 2000 and 2003, with a total of 47 OECD and non-OECD countries participating. The 2006 cycle focuses on scientific literacy. PISA is distinguished by its focus on skills for life. Its subject matter is grounded in the broader concept of literacy, or the ability to use and apply knowledge.

- The Adult Literacy and Lifeskills (ALL) Study is coordinated by Statistics Canada, in conjunction with OECD. ALL is a large-scale comparative survey designed to measure a broad range of skills in the adult population aged 16–65. The skills are considered important for social and economic success and continuous learning, and include prose and document literacy, numeracy, and problem solving. Builds on the International Adult Literacy Study (IALS), ALL is a household survey being administered in 2003 and 2005. So far, four non-OECD countries have indicated their intention to participate.

Additionally, many non-OECD and developing countries are establishing or joining networks focused on developing systems for collecting comparable educational data, including learning outcomes measurement. Examples of two key regional initiatives follow:

- Under UNESCO auspices, the Latin American Educational Quality Assessment Laboratory was founded in 1994 to field a comparative regional assessment of the quality of educational outcomes. Eleven countries participated in the 1997 assessment of 3rd- and 4th-grade students in language and mathematics (Casassus et al. 1998). This was the first comparative assessment experience for several participating countries.

- The South African Consortium for Monitoring Educational Quality (SACMEQ) is a voluntary network of 15 ministries of education in eastern and southern Africa. SACMEQ was established in 1995 for the purpose of undertaking joint training and policy research related to education. Its mission is to equip educational planners in member countries with the technical skills needed for effectively monitoring and evaluating schooling and the quality of education. Core activities include collecting data on student outcomes in reading and mathematics and providing technical support and capacity building in monitoring and evaluation systems, including assessment delivery.

Websites that provide more information on these and selected national activities are listed in the box below.

Websites Relating to International and Regional Activities

International and Regional Activities	
PIRLS and TIMSS	http://www.timss.org
PISA	http://www.pisa.oecd.org
ALL	http://www.ets.org/all
Latin American Educational Quality Assessment Laboratory	http://www.unesco.cl/09.htm (Spanish)
SACMEQ	http://www.unesco.org/iiep/eng/networks/sacmeq/sacmeq.htm
Select National Activities	
Brazil's SAEB	http://www.inep.gov.br/idiomas/ingles/
Canada's SAIP	http://www.cmec.ca/saip/indexe.stm
England's National Curriculum Assessment	http://www.qca.org.uk
New Zealand's NEMP	http://nemp.otago.ac.nz/
United States' NAEP	http://www.nces.ed.gov/nationsreportcard/

Initiating Measurements of System-Level Learning Outcomes

In developing countries, the measurement of system-level learning outcomes presents unique challenges. Building and administering a system-level assessment of student outcomes is a massive endeavor. It requires a clear timeline, substantial human and financial resources, and a wide range of expertise. Some of this expertise may be available from ministry staff or personnel in other levels of the education system. Some expertise may be hired through local or international consultants or organizations that help build capacity for assessment-related activities. Developing country participation in international assessments is very helpful, allowing researchers to network, observe assessment functions and phases, build skills, and realize macro-level data.

At the outset during the development of the overall strategy, it is important to involve visionaries at the policy level: those who see the big picture and can provide clear direction about why the assessment program is needed. At this point it is also important to involve strategic planners who can help stakeholders think through the implications of design decisions. During development and implementation, the effort will require other areas of expertise that, ideally, are not mutually exclusive. These include

- subject matter or curriculum experts to provide guidance on constructing test instruments and review the relevance and appropriateness of items;
- assessment and survey specialists to coordinate the effort and ensure that technical standards are met;
- psychometricians to provide guidance on technical aspects of constructing the test and scaling the data; and
- other statisticians and analysts to construct the database, analyze data, and assist in report production

The financial costs of developing and implementing a high-quality assessment program depend on its scope, the size of the system assessed, and how the effort is staffed. For example, SAIP costs approximately CAN$3 million annually. It assesses three subject areas, each for two age groups, over four years, and provides national, provincial, and school-level data. The Government of Canada contributes about half the costs to the contractor developing and managing the program. Provincial authorities supply the other half. Assessment costs may be less in other countries. For example, in Portugal assessment development and management are not contracted out; education ministry staff are responsible.

Developing the overall strategy can take several years. However, it is an important process. Time invested in developing consensus around key purposes, design features, and desired assessment products should pay off in wide buy-ins and smooth implementation and reception. Once the strategy is developed, a timeline for development, field testing, and main data collection can be set. An ambitious one may be little as three years. The cycle established could be shortened once the program is up and running. Table 2 presents a sample timeline and major tasks in fielding a system-level assessment for the first time.

Table 2. Sample Timeline and Tasks for Developing and Fielding an Assessment (adapted from PISA data strategy)

Timing	Tasks
Starting point	Overall strategy for assessment developed in conjunction with policymakers and key stakeholders.
Year 1	• Experts and consultants identified by staff. • Frameworks and test specifications developed. • Test material collected or developed. • Test assembled after initial item review. • Sampling plan and draw sample developed. • Context questionnaires drafted.
Year 2	• Field-test manuals prepared. • Procedures developed for test receipt and data entry. • Cooperation obtained for field test; promotional materials developed. • Materials translated and verified, if necessary. • Instruments and questionnaires field tested. • Results analyzed and revisions made to instruments and processes (manuals). • Sample drawn for main data collection and effort begun to obtain cooperation from schools.
Year 3	• Final instruments and questionnaires translated and verified. • Quality control training conducted. • Marker training conducted. • Effort continued to obtain cooperation for main data collection, • Data collected. • Marking and data entry completed; database constructed. • Sampling weights computed and scaling undertaken. • Summary tabulations prepared.
Year 4	• Analyses undertaken. • Technical documentation prepared; reports produced.
Throughout	• Groups convened as necessary.

Considerations for the Future

An issue requiring special consideration in developing countries is the design approach taken to domains. As noted previously, assessment designers can focus on curriculum or on broad concepts such as literacy. The first is straightforward, and ample material is likely to be available for test development. Further, relating assessment results to what happens in classrooms is a potentially easier task. However, the assessment will reflect any limitations of the curriculum, which, in any case, may not manifest what assessors are trying to learn about the education system. Further, to the extent that curricular variation exists, common elements used to fairly assess students may not reflect measurements desired. Moreover, curriculum achievement may be verified through examinations, classroom-level activities, or in other ways.

An approach identifying what should be learned—perhaps for personal development or economic success—may better capture assessment objectives, though it will be more difficult to explain or relate to teaching practice. For example, the lifeskills approach may tap into a broader range of knowledge or skills that students develop outside of school. To the extent that assessments can drive what is taught and counter the phenomenon of "teaching to the test," this approach might positively influence education practice in new directions.

A related, forward-looking question is whether domains chosen for measurement are limited to strict subject areas. Increasingly, teachers are encouraged to teach subjects in integrated ways that are more reflective of real-life situations. Future assessments might focus on integrated subject areas, though careful thought has not been given to interpreting and making use of their data.

In many countries, affective outcomes— positive dispositions for learning, motivation, self-assurance, and flexibility—are attributes schools try to impart. The measurement of affective outcomes—whether directly through assessment or indirectly through questionnaires—may be an important component of assessment programs.

The issue of engagement in school and motivation for learning may be of particular interest for developing countries, where improving the quality of teachers and teaching remains the primary vehicle for improving education. Learning how teachers influence students' motivation may provide valuable information and help improve outcomes. In OECD countries, this issue is being explored as a key intermediary link between teaching and achievement.

The sustainability of any assessment program relies on its trustworthiness and usefulness. Ultimately, these qualities depend on the care with which information is interpreted and presented. Rushing to judgment can undermine the public's trust, and conservative approaches to analysis and explanation of data are needed. The proper course is to look carefully for explanations and possible measurement errors, then tailor assessment results to different audiences, providing the kind of information each requires to effect positive change.

The measurement of affective outcomes—whether directly through assessment or indirectly through questionnaires—may be an important component of assessment programs.

Acknowledgements

This paper was written for EQUIP2 by Maria Stephens and Jay Moskowitz (AIR), 2004.

References

Bolger, N., and T. Kellaghan. 1990. "Method of Measurement and Gender Differences in Scholastic Achievement." *Journal of Educational Measurement* 27: 165-174.

Casassus, Juan. 1999. *First International Comparative Study of Language, Mathematics, and Associated Factors in Third or Fourth Grade*. UNESCO. The Major Project of Education, Bulletin 50, December 1999. <http://www.unesco.cl/pdfactyeven/ppe/boletin/artingle/50-3.pdf>

Guimarães de Castro, Maria Helena. 2001 *Education Assessment and Information Systems in Brazil*. Brasilia: National Institute for Educational Studies and Research.
Kirsch, Irwin. 2003. "The PISA Framework for Assessing ICT Literacy." Paper prepared for OECD/INES/Network A, April 2003.

Organisation for Economic Cooperation and Development. 1999. *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Paris: OECD. http://www.pisa.oecd.org/Docs/download/PISAFrameworkEng.pdf

UNESCO. 2000. *Dakar Framework for Action. Education for All: Meeting Our Collective Commitments. Adopted by the World Education Forum, Dakar, Senegal, April 26–28, 1999.* Paris: UNESCO. http://unesdoc.unesco.org/images/0012/001211/121147e.pdf

UNESCO. 1990. "World Declaration on Education for All: Meeting Basic Learning Needs." Statement adopted by the World Conference on Education for All, Jomtien, Thailand, March 1990. <http://www.unesco.org/education/efa/ed_for_all/background/jomtien_declaration.shtml>

Appendix 1. Sample Item Types

The following sample items demonstrate three item types, across grades and subject areas. The wording of the items is presented verbatim but their formatting has been modified.

Sample multiple-choice item from TIMSS Grade 4 science assessment

Which one of the following characteristics is most likely to be found in mammals that are preyed on by other mammals?

a. Eyes on the side of the head
b. Teeth that are long and pointed
c. Claws on the feet
d. Ears that cannot move

[Correct answer: a; 37 percent of students answered correctly.]

Sample short constructed-response item from New Zealand's NEMP mathematics assessment for years 4 and 8

This picture shows a busy roadway. During the day time, about 98 cars go down this road every minute. About how many cars would go down this road in 9 minutes?

[Correct answers: 882 or 900. 48 percent of year 8 students answered this item correctly; 28 percent used estimation to arrive at the answer.]

Sample extended constructed-response item from the U.S. NAEP Grade 4 civics assessment

In Ms. Tanaka's fourth-grade class, students must decide which one of three books will be read aloud. Ms. Tanaka agreed to let the students decide whether they will all read *Green Fields, The Lion That Saved My Life,* or *The Wanderers.* She told them it is up to them to choose, but they must do so in a fair and democratic way.

Of the 25 students in Ms. Tanaka's class, three—Wanda, Marcello, and Ellis—have already read *Green Fields.* Both Ellis and Wanda have also read *The Lion That Saved My Life*. No one in the class has read *The Wanderers.*

What would be the most democratic way of selecting a book to read aloud in class? Why would this be the most democratic way?

[Correct response describes and justifies a democratic process: voting; selecting representatives; fairness; and one person, one vote. Seven percent of students received full credit and 35 percent received partial credit.]

Appendix 2. Non-OECD Countries Participating in International and Select Regional Assessments

Country	*PIRLS 01*	*TIMSS 03*	*PISA 03*	*ALL*	*SACMEQ*	*Laboratory*
Argentina	X	X				X
Armenia		X				
Bahrain		X				
Belize	X			X		
Bermuda				X		
Bolivia				X		X
Botswana		X			X	
Brazil			X			X
Bulgaria	X	X				
Chile		X				X
China						
Chinese Taipei		X				
Colombia	X					X
Costa Rica				X		X
Cuba						X
Cyprus	X	X				
Dominican Republic						X
Egypt		X				
Estonia		X				
Ghana		X				
Honduras						X
Hong Kong	X	X	X			
Indonesia		X	X			
Iran	X	X				
Israel	X	X				
Jordan		X				
Kuwait	X	X				
Latvia	X	X	X			
Lebanon		X				
Lesotho					X	
Lithuania	X	X				
Macedonia	X	X				
Malaysia		X				
Moldova	X	X				
Morocco	X	X				
Mozambique					X	
Paraguay						X
Peru			X			X
Philippines		X				
Romania	X	X				
Russian Federation	X	X	X			
Saudi Arabia		X				
Serbia		X	X			
Seychelles					X	
Singapore	X	X				
Slovenia	X	X				
South Africa		X			X	
Swaziland					X	
Syria		X				
Tanzania					X	
Thailand			X			
Tunisia		X	X			
Uganda					X	
Uruguay			X			
Venezuela						X
Yemen		X				

Appendix 3. How Learning Outcomes Are Measured in Brazil

Although there is little systematic knowledge about how learning outcomes are measured in non-OECD countries, some countries are developing or have developed national-level assessment programs, including Brazil, Colombia, Egypt, Jordan, Namibia, and Thailand. Within this field, Brazil is considered a leader.

The National Basic Education Evaluation System (SAEB), underway in Brazil since 1990, is an example of low-stakes measurement of learning outcomes at the system level. SAEB assesses a sample of 4th-, 8th-, and 11th-grade students across Brazil's 26 states and the federal district. Key subjects, such as Portuguese language and mathematics, are assessed every two years. As such, the assessment is a cross-sectional design. Results are reported in terms of percentages of students reaching successive proficiency levels.

SAEB's goal is to identify weaknesses in the educational system and factors related to high and low performance. From past assessments, Brazilian policymakers learned that factors amenable to reform that account for achievement differences include gaps between the proposed and learned curriculum and differences in teachers' qualifications (Guimarães de Castro 2001).

Other low-stakes system-level assessments include state-level programs, such as those implemented in Bahia, and participation in international activities, including OECD's Program for International Student Assessment and UNESCO's Latin American Educational Quality Assessment Laboratory.

Other types of learning outcomes measurement exist in Brazil. The National Secondary Education Examination (ENEM) measures students' individual performance after basic education, checking competencies and abilities that are essential for active citizenship and related to a new curriculum. ENEM is voluntary, but it can be categorized as a high-stakes measurement of individual outcomes since it may be used as an alternative to the university entrance exam.

The National Course Examination (or Provão), is mandatory for university students completing select courses. Although individual grades are provided, they are confidential, used in the aggregate to assess the overall quality of the undergraduate courses.

Activities to measure learning outcomes are complemented by more traditional activities to measure administrative aspects of the education system. These include annual school and higher education censuses, which, together with outcome data, are part of Brazil's comprehensive monitoring and evaluation system.

EQUIP2: Educational Policy, Systems Development, and Management is one of three USAID-funded Leader with Associate Cooperative Agreements under the umbrella heading Educational Quality Improvement Program (EQUIP). As a Leader with Associates mechanism, EQUIP2 accommodates buy-in awards from USAID bureaus and missions to support the goal of building educational quality in the national, sub-national, and cross-community levels.

The Academy for Educational Development (AED) is the lead organization for the global EQUIP2 partnership of education and development organizations, universities, and research institutions. The partnership includes fifteen major organizations and an expanding network of regional and national associates throughout the world: Aga Khan Foundation USA, American Institutes for Research, CARE USA, East West Center, Education Development Center, International Rescue Committee, Joseph P. Kennedy, Jr. Foundation, Learning Communities Network, Michigan State University, Mississippi Consortium for International Development, ORC Macro, Research Triangle Institute, University of Minnesota, Institute for International Studies in Education at the University of Pittsburgh, Women's Commission for Refugee Women and Children.

For more information about EQUIP2, please contact

USAID
Patrick Collins
CTO EGAT/ED
USAID Washington
1300 Pennsylvania Ave., NW
Washington, DC 20532
Tel: 202-712-4151
Email: pcollins@usaid.gov

AED
John Gillies
EQUIP2 Project Director
1825 Connecticut Ave., NW
Washington, DC 20009
Tel: 202-884-8256
Email: EQUIP2@aed.org
Website: www.equip123.net

EQUIP2 is funded by the United States Agency for International Development
Cooperative Agreement No. GDG-A-00-03-00008-00